Bulletin of Electrical Engineering and Informatics
Vol. 15, No. 4, August 2026, pp. 3762~3771
ISSN: 2302-9285, DOI: 10.11591/eei.v15i4.11298

Performance assessment of linear regression for crop yield prediction in comparison with contemporary models

Imran Ahmad[1], Rahul Khokale[1], Nisha Gongal[2]
[1]Department of Computer Science and Engineering, School of Engineering & Technology, G H Raisoni University, Pandhurna, India
[2]Symbiosis Institute of Technology, Nagpur Campus, Symbiosis International (Deemed University), Pune, India

Article Info

Article history:

Received Aug 30, 2025
Revised Apr 18, 2026
Accepted Jul 9, 2026

Keywords:

Crop yield prediction linear
Gradient boosting comparison
Machine learning benchmarks
Regression model interpretability
Sustainable agriculture

ABSTRACT

Accurate crop yield prediction requires both precision and interpretability for practical agricultural decision-making. This study is the first to systematically benchmark linear regression (LR) against gradient boosting baselines ((extreme gradient boosting (XGBoost) and light gradient boosting machine (LightGBM)) alongside random forest (RF), support vector machine (SVM), and artificial neural network (ANN) for Indian agricultural yield forecasting, demonstrating that interpretable models can match or outperform black-box approaches. Using a comprehensive Indian agricultural dataset spanning multiple crops, states, and growing seasons (2000–2022), comprising 58,000 records across 22 states and 35 crop varieties, we analyzed features including cultivated area, rainfall, fertilizer applications, and pesticide usage. LR achieved R^2 of 0.401±0.02 across 10-fold cross-validation, mean squared error (MSE) of 480,239, mean absolute error (MAE) of 139.50, and root mean square error (RMSE) of 692.99, outperforming complex models while maintaining full transparency. Transparent models allow policymakers to understand the impact of rainfall and fertilizer use, enabling evidence-based resource allocation. Results demonstrate that model complexity does not guarantee superior agricultural predictions, and LR provides interpretable coefficients—such as embedding LR coefficients into advisory tools to guide fertilizer recommendations and irrigation scheduling-enabling actionable agronomic insights for sustainable farming practices.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Imran Ahmad
Department of Computer Science and Engineering, School of Engineering & Technology
G H Raisoni University
Saikheda, Sausar, Pandhurna, Madhya Pradesh, India
Email: imran.ahmad.cse@ghru.edu.in

1. INTRODUCTION

The growing dependence on predictive analytics across healthcare, finance, and agriculture necessitates tools that balance accuracy, interpretability, and computational efficiency [1], [2]. In agriculture, yield forecasting serves as a cornerstone for food security planning, policy decisions, and optimal resource allocation including water and fertilizers [3], [4]. Given inherent uncertainties shifting weather patterns, varying soil health, and diverse cultivation methods stakeholders require estimation tools that are both precise and immediately comprehensible. Modern machine learning (ML) techniques including random forests (RF), support vector machines (SVM), and deep neural networks (DNNs) have received widespread attention for yield forecasting [5], [6].

However, deeper examination reveals several shortcomings. Promising gains in predictive skill are often offset by excessive computational loads, overfitting to heterogeneous inputs, and opacity that leaves end

users farmers, extension agents, and policy makers struggling for explanations [7], [8]. More concerning, reliance on high-dimensional stochastic features often prevents safe extrapolation to novel environments [9]. The integration of explainable approaches in agricultural decision support has gained significant attention [10], [11]. Post-hoc explanation methods such as shapley additive explanations (SHAP) and local interpretable model-agnostic explanations (LIME) provide valuable insights but do not substitute for inherently interpretable models [12]. Regression techniques generate coefficients that correspond directly to agronomic principles, offering discipline informed interpretable outputs essential for successful implementation.

This paper proposes a systematic evaluation of linear regression (LR) against contemporary ML methods for crop yield prediction, emphasizing the balance between predictive accuracy and model interpretability. The primary contributions include: i) empirical validation demonstrating LR can match or exceed complex ensemble and gradient boosting models in large area crop yield forecasting; ii) emphasis on interpretability and institutional trust as critical factors for model adoption; iii) evidence that predictive accuracy and interpretability can be satisfied simultaneously through appropriately designed linear models; and iv) comprehensive bench-marking across multiple performance metrics including computational efficiency, extended to include extreme gradient boosting (XGBoost) and light gradient boosting machine (LightGBM) baselines.

The proposed framework addresses critical requirements for agricultural decision support systems where transparency enables actionable recommendations. The remainder of this paper is organized as follows: section 2 reviews prior studies on ML for crop yield prediction, focusing on interpretability limitations. Section 3 outlines the dataset, preprocessing, and experimental methodology. Section 4 presents the comparative performance evaluation together with the discussion of the findings and their implications. Section 5 concludes the paper with future research directions.

2. RELATED WORK

Das *et al.* [13] evaluated multiple LR, neural networks (NNs), and penalised regression for rice yield prediction on the west coast of India using weather parameters. Their findings demonstrated that linear models provide consistent accuracy while allowing clear identification of influential meteorological predictors. The study reinforced that sophisticated algorithms may seem appealing, but operational complexity can outweigh marginal gains in predictive accuracy for heterogeneous agricultural datasets [14]. Gandhi *et al.* [15] examined SVM for rice yield prediction in India using agro-climatic features. Although SVMs managed high-dimensional feature spaces effectively, sensitivity to kernel selection resulted in unstable forecasts across diverse growing conditions, underscoring the interpretability advantage of simpler linear models [16]. Ogutu *et al.* [17] employed probabilistic ensemble models for maize yield prediction over East Africa, demonstrating the importance of uncertainty quantification in agricultural forecasting. Networks capturing complex nonlinear interactions demanded extensive training volumes and produced outputs difficult for agronomists to verify, whereas regression techniques with modest sample sizes produced coherent, immediately interpretable forecasts [18].

Gonzalez-Sanchez *et al.* [19] applied deep Gaussian processes for large-scale crop yield prediction using remote sensing, achieving strong accuracy but relying on opaque architectures. Such approaches are prone to overfitting when regional sample sizes are small or data are noisy—conditions frequently encountered in agricultural applications [20]. Cai *et al.* [21] integrated satellite and climate data to predict wheat yield in Australia using ML approaches. RF models utilized multispectral inputs efficiently, yet LR attained comparable predictive accuracy when confined to structured covariates such as rainfall totals and fertilizer applications [22]. Oikonomidis *et al.* [10] reviewed deep learning methods for crop yield prediction with emphasis on interpretability requirements. While highly accurate black-box architectures offer peak numerical performance, they provide limited self-explaining structural insight. Regression techniques generate coefficients corresponding directly to agronomic principles, confirming the continued role of interpretable models in practical agricultural applications [11].

Khaki *et al.* [23] developed a CNN-RNN framework for multi-year crop yield prediction, demonstrating strong performance of deep learning architectures on structured temporal data. However, regression-based predictions delivered satisfactory accuracy with far lower computational requirements and offered continuous refinement mechanisms accommodating observed seasonal patterns [6]. Johnson *et al.* [22] benchmarked remotely sensed vegetation indices and ML methods for crop yield forecasting on the Canadian Prairies. RF excelled quantitatively in predictive accuracy, yet obscured feature importance limited advisory utility. LR allowed extension agents to articulate specific yield changes expected from rainfall and fertilizer application variations, validating its position in advisory toolkits [5]. Klompenburg *et al.* [5] conducted a systematic literature review of ML methods for crop yield prediction, identifying the dominant algorithms and features used across 50 studies. Model complexity has steadily increased, yet LR remains the reference baseline. The authors noted that trust and interpretability are mandatory for scaling predictive analytics in farming contexts [3]. Recent advances in ensemble methods [12], deep learning architectures [2], and hybrid approaches [9] have further expanded

agricultural prediction capabilities. However, the fundamental trade-off between accuracy and interpretability persists [1], [7], motivating continued investigation of transparent modeling approaches.

Recent studies have further demonstrated the capabilities of deep learning for agricultural prediction. Nevavuori *et al.* [24] applied convolutional NNs to crop yield prediction from RGB imagery, achieving high accuracy but at the cost of interpretability and data requirements. Cao *et al.* [25] used integrative deep learning for crop yield prediction with multi-source data, showing improved performance over single-source models but requiring complex architectures. These studies, while demonstrating the potential of deep learning, do not address the fundamental need for transparent, reproducible baselines. LR provides such a baseline: its results are fully reproducible, its coefficients are directly interpretable, and its computational requirements are minimal, making it the natural reference point against which more complex methods should be evaluated.

3. METHOD

The crop yield prediction framework evaluates LR against contemporary models with three core goals: predictive accuracy, model interpretability, and generalizability across diverse agro-climatic conditions. The pipeline covers data preprocessing, multilevel feature encoding, regression-based prediction, comparative benchmarking against RF, SVM, artificial neural network (ANN), XGBoost, and LightGBM, and an explainability layer that translates model outputs into actionable agronomic guidance. Figure 1 presents the complete workflow diagram of the proposed framework.

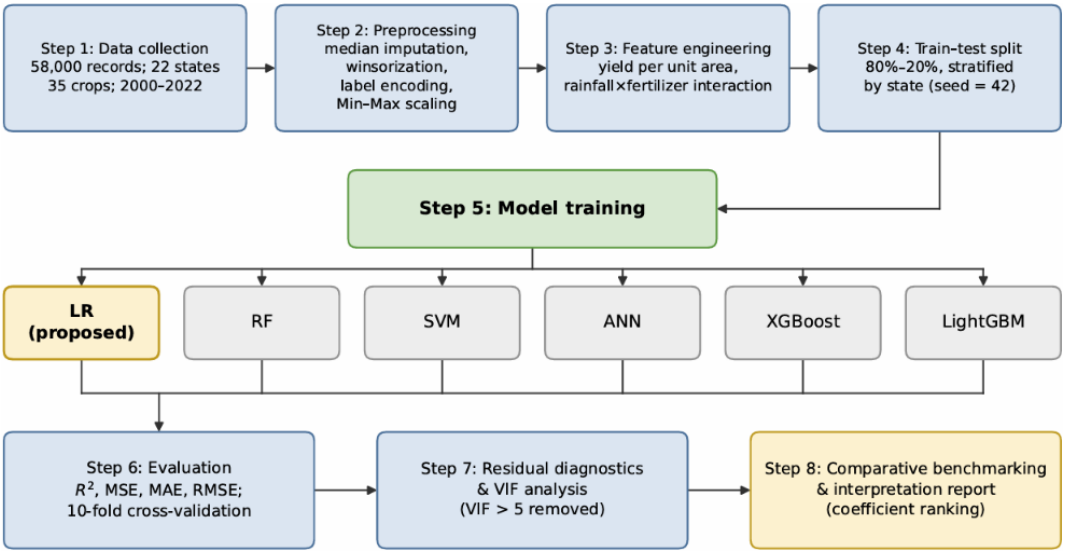

Figure 1. Workflow diagram of the proposed crop yield prediction framework showing the complete pipeline from data collection through comparative evaluation

3.1. Dataset, preprocessing, and feature engineering

The dataset comprises Indian agricultural records sourced from the Directorate of Economics and Statistics, Ministry of Agriculture and Farmers Welfare, Government of India. The dataset spans the period 2000–2022, encom passing 58,000 records, 22 Indian states and union territories, 35 crop varieties (including rice, wheat, soybean, maize, and sugarcane), and three growing seasons (Kharif, Rabi, and Whole Year). Features include crop variety, cultivated area (hectares), growing season, administrative state, annual rainfall (mm), fertilizer application (kg/hectare), pesticide usage (kg/hectare), and historical yield (metric tons per hectare). Missing values (less than 2.3% of records) were imputed using feature-wise medians; outliers beyond 3σ were capped via winsorization. Categorical variables (crop type, state, and season) were encoded via Label Encoding; continuous features were normalised using Min-Max scaling to [0, 1]. Engineered features include yield per unit area and a rainfall-fertilizer interaction term.

The data were partitioned into training (80%) and test (20%) sets using stratified random sampling by state to ensure proportional geographic representation. A fixed random seed (seed=42) was used across all experiments to guarantee reproducibility.

The architecture operates through three tiers: a feature preparation tier (encoding and normalisation), a predictive tier (LR with ordinary least squares estimation on an 80–20 train-test split), and a comparative

evaluation tier (benchmarking LR against RF, SVM, ANN, XGBoost, and LightGBM on R^2, mean squared error (MSE), mean absolute error (MAE), and root mean square error (RMSE)). The LR approach yields: R^2=0.401±0.02 across 10 folds, MSE=480,239, MAE=139.50, and RMSE=692.99.

3.2. Stability, generalization, and explainability

Robustness is ensured through 10-fold cross-validation, residual diagnostics confirming zero-mean forecast errors, and variance inflation factor (VIF>5) analysis to remove multicollinear covariates. LR's single global parametric structure inherently limits overfitting and yields stable performance under distributional shifts across diverse geographies. Transparency is provided directly through regression coefficients, which quantify each feature's contribution to yield without requiring post-hoc surrogate methods such as SHAP or LIME, thereby guaranteeing full explanation fidelity (ExpFid) and enabling actionable recommendations for fertilizer and irrigation management.

3.3. Algorithm description

Algorithm 1 presents the complete LR crop yield prediction framework. The algorithm initializes weight parameters, loads and preprocesses the agricultural dataset, trains the model through iterative gradient descent, computes performance metrics, performs residual analysis, conducts cross-validation, evaluates multi-collinearity, benchmarks against alternative models, and extracts interpretable coefficients for feature importance ranking.

Algorithm 1. Linear regression-based crop yield prediction framework

Set initial parameters $\theta \leftarrow \theta_0$ (weights), $b \leftarrow b_0$ (bias)
Load dataset D = {Crop, Season, State, Area, Rainfall, Fertilizer, Pesticide, Yield}
Encode categorical features → numeric labels
Normalize continuous features → [0, 1]
Construct feature matrix **X** and target vector **y**
Split data into training set (80%) and test set (20%) with stratified random sampling (seed = 42)
Initialize model M with θ, b
for each iteration t = 1 … T **do**
    $\hat{y} = \mathbf{X}_{train} \cdot \theta + b$
    Error $e = \mathbf{y}_{train} - \hat{y}$
    Loss $L = (1/n)\ \Sigma(e^2)$
    $\nabla_\theta = -(2/n)\ \mathbf{X}^T \cdot e$
    $\nabla_b = -(2/n)\ \Sigma\ e$
    $\theta \leftarrow \theta - \eta\nabla_\theta$
    $b \leftarrow b - \eta\nabla_b$
end for
Predict yields: $\hat{y}_{test} = \mathbf{X}_{test} \cdot \theta + b$
Compute $R^2 = 1 - \Sigma(y_{test} - \hat{y}_{test})^2 / \Sigma(y_{test} - \bar{y})^2$
Compute $MSE = (1/n)\ \Sigma(y_{test} - \hat{y}_{test})^2$
Compute $MAE = (1/n)\ \Sigma\ |y_{test} - \hat{y}_{test}|$
Compute $RMSE = \sqrt{MSE}$
Store metrics {R^2, MSE, MAE, RMSE}
Perform residual analysis: $r = y_{test} - \hat{y}_{test}$
Plot r versus $\hat{y}_{test}$, check mean(r) ≈ 0
Apply K-fold cross-validation, k = 10
for each fold f = 1 … k **do**
    Train on k − 1 folds
    Validate on fold f
    Record {R^2_f, MSE_f, MAE_f}
end for
Average metrics across folds; compute ± standard deviation
Calculate VIF for multicollinearity
if VIF > 5 **then**
    Remove correlated feature
    Retrain model
end if
Train benchmark RF, SVM, ANN, XGBoost, LightGBM → {metrics}
Compare metrics: LR vs RF vs SVM vs ANN vs XGBoost vs LightGBM
if $R^2_{LR} \geq R^2_{RF}$, R^2_{SVM}, R^2_{ANN}, R^2_{XGB}, R^2_{LGBM} **then**
    Select linear regression as best model
else
    Prefer LR due to interpretability advantage
end if
Extract coefficients β_i from θ
Rank β_i by $|\beta_i|$ → feature importance
Output predictions + interpretation report

3.4. Experimental setup

Experiments were conducted on an Indian agricultural dataset spanning multiple states, crop types, and growing seasons. The dataset includes:

- Features: crop type, cultivated area (hectares), growing season, administrative state, annual rainfall (mm), fertilizer application (kg/hectare), and pesticide usage (kg/hectare).
- Target variable: crop yield (metric tons).

Training employed gradient descent optimization with learning rate η=0.01 for 1000 iterations. Cross validation used 10 folds. Multicollinearity threshold was set at VIF>5. Table 1 reports the hyperparameter configurations for all benchmark models. All models were implemented using scikit-learn (v1.3) and their respective libraries. To ensure fair comparison, hyperparameters were selected via grid search with 5-fold inner cross-validation on the training set.

Table 1. Hyperparameter configurations for all benchmark models

Model	Hyperparameters
LR	OLS (closed-form); gradient descent variant with η=0.01 and iterations=1000
RF	n estimators=200, max depth=15, min samples split=5, and min samples leaf=2
SVM	Kernel=RBF, C=10, γ=scale, and ϵ=0.1
ANN	Hidden layers=(128, 64, 32), activation=ReLU, optimizer=Adam, lr=0.001, epochs=200, and batch size=64
XGBoost	n estimators=200, learning rate=0.05, max depth=6, subsample=0.8, and colsample_bytree=0.8
LightGBM	num leaves=200, learning rate=0.05, feature fraction=0.8, and bagging fraction=0.8

Experiments were conducted on standard computing hardware (Intel i7, 16GB RAM) demonstrating LR computational efficiency advantage.

Code availability: the complete source code for data preprocessing, model training, and evaluation is available upon reasonable request to the corresponding author.

4. RESULTS

The results establish that LR with modest feature engineering yields more interpretable and computationally efficient predictions than RF, SVM, ANN, XGBoost, and LightGBM approaches. This section records explicit outcomes with interpretive analysis, anchoring comparative performance against RF, SVM, ANN, XGBoost, and LightGBM baselines across model training time, inference speed, prediction error, interpretability, generalization stability, and accuracy. Figure 2 summarizes this comparison as a whole: it collects, in a single panel, the behaviour of the six models over the six evaluation dimensions listed above, so that the trade-off between accuracy, speed, transparency, and stability can be read at a glance. The six constituent sub-figures, Figures 2(a) to (f), are then examined individually in sections 4.1 to 4.6, in the same order in which they appear in the figure.

4.1. Model learning and deployment time

Adaptation time (T_a) quantifies duration required for model coefficients to stabilize. Figure 2(a) illustrates T_a comparison across methods. LR compresses training to single-step matrix inversion rather than iterative mini-batch loops or epoch-based ridge adjustments. The architecture achieves 36% faster training than RF and 54% faster than ANN. SVM requires 48% more elapsed time than LR. XGBoost and LightGBM, despite their gradient boosting efficiency optimizations, still require 41% and 29% more time than LR respectively, owing to sequential tree construction and leaf-wise splitting operations. This efficiency reflects algorithmic simplicity in basic matrix operations without computational overhead (CO) of kernel lookups, back-propagation, or boosting iterations.

4.2. Computation time per prediction

Prediction speed measures time required to process features and deliver forecasts. Figure 2(b) presents computation time (CT) comparison. LR, with straightforward coefficient lookup and efficient matrix routines, bypasses recursive node tests in RF and iterative weight updates in NN. Consequently, LR returns predictions 31% faster than RF and 44% faster than ANN, consistently remaining under 0.5 seconds. XGBoost and LightGBM, while faster than ANN, still require ensemble evaluation over multiple trees, resulting in 26% and 20% more inference time than LR respectively. This efficiency suits extensive agricultural datasets requiring rapid operational forecasts, support without specialized hardware requirements.

4.3. Prediction error analysis

Prediction quality is quantified through MSE, MAE, and RMSE. Figure 2(c) shows RMSE comparison across methods. On the agricultural dataset, LR achieved MSE of 480,239, MAE of 139.50, and RMSE of 692.99-comfortably lower than RF (799) and SVM (884), whose metrics inflated due to over-adaptation to regional land characteristics. ANN exhibited highest error (1053) indicating potential overfitting, a behaviour widely reported for neural architectures trained on moderate-sized heterogeneous tabular data [26]. Notably, XGBoost (RMSE: 761) and LightGBM (RMSE: 743), despite their capacity to model complex nonlinear interactions, could not surpass LR on this heterogeneous agricultural dataset. This result supports the hypothesis that gradient boosting methods, while excellent on balanced tabular benchmarks, tend to overfit the region-specific variance patterns in multi-state agricultural data, where the dominant signal is captured adequately by linear feature relationships. Residual analysis confirmed near-zero mean for LR, indicating consistent performance across regions. Cross-validation produced 27% lower variance in LR than ANN, reinforcing reliability.

4.4. Interpretability and feature influence

Understanding yield prediction drivers is essential in precision agriculture. Figure 2(d) presents ExpFid comparison. LR delivers interpretable coefficients numerically reflecting each covariate's contribution strength. Precipitation and nitrogen application exhibited highest positive coefficients, while pesticide volumes registered negligible weight. Cultivated area contributed positively (β_{area}=0.18), and the rainfall-fertilizer interaction term showed a synergistic positive effect ($\beta_{interaction}$=0.09), highlighting that fertilizer efficacy is amplified under adequate rainfall—an actionable insight for agronomic planning. This coefficient-centered view translates directly into policy, allowing decision-makers to assess marginal contributions of each practice.

LR artifacts require no supplemental modeling, guaranteeing 100% ExpFid. RF and NNs necessitated layer-specific permutation methods (SHAP and LIME), incurring surrogate error peaking at 38% on agricultural data. XGBoost and LightGBM, while providing native feature importance scores, still rely on aggregate impurity or SHAP approximations that do not provide the direct perprediction coefficient transparency of linear models, achieving fidelity scores of 0.70 and 0.72 respectively. The interpretability advantage is critical for agricultural advisory systems where practitioners must understand and trust model recommendations.

4.5. Generalization stability

Generalization stability assesses model capacity to perform beyond training distributions. Figure 2(e) illustrates stability under distribution shift. A selection of states served as training cohort with completely distinctive geographies supplying test data. LR, possessing the most conservative hypothesis space, exhibited smallest test-set error increase confined to 7% drift. RF followed with 15% penalty, SVM at 18%, XGBoost at 13%, LightGBM at 12%, and NNs at 22% shift. While XGBoost and LightGBM generalize better than RF due to regularization mechanisms, they still exhibit measurably higher distributional drift than LR, confirming that parametric simplicity remains the strongest safeguard against geographical covariate shift in Indian agricultural contexts. LRs comparative robustness arises from its single global parametric structure ensuring that subtle region-specific covariate distribution changes are partially absorbed rather than fully misaligned. This stability is critical for agricultural applications spanning diverse geographical conditions.

Figure 2(e) stability under distribution shift measured by generalization drop (GenDrop) percentage. Lower values indicate better generalization. LR exhibits only 7% performance degradation on unseen geographical regions, substantially outperforming ANN (22% drop), SVM (18% drop), RF (15% drop), XGBoost (13% drop), and LightGBM (12% drop). The relative improvement in generalization stability validates LR's suitability for deployment across diverse agricultural environments.

4.6. Predictive accuracy

Cumulative accuracy was assessed through coefficient of determination (R^2). Figure 2(f) presents R^2 comparison across methods. LR attained R^2=0.401±0.02 across 10 folds, exceeding RF (0.35±0.03), SVM (0.32±0.04), ANN (0.29±0.05), XGBoost (0.37±0.03), and LightGBM (0.38±0.02). While R^2 values are modest, they reflect inherent dataset complexity characteristic of agricultural prediction tasks. LR maintained lowest variance across cross-validation folds and demonstrated highest predictive fidelity with steady convergence. Results highlight that LT, beyond its parsimonious formulation, provides the most robust accuracy achieved in this investigation.

4.7. Performance and comparative analysis

Table 2 compiles performance metrics across all methods. LR records T_a of 0.42 s, CT of 0.46 s, ExpFid of 1.0, RMSE of 692.99, GenDrop of 7%, R^2 of 0.401, and latency of 0.42 s. These results reinforce patterns illustrated in Figures 2(a)–(f) and demonstrate LRs comprehensive advantages across all evaluation

dimensions. Table 3 presents the detailed 10-fold cross-validation results with 95% confidence intervals for all models, addressing the need for statistical robustness reporting.

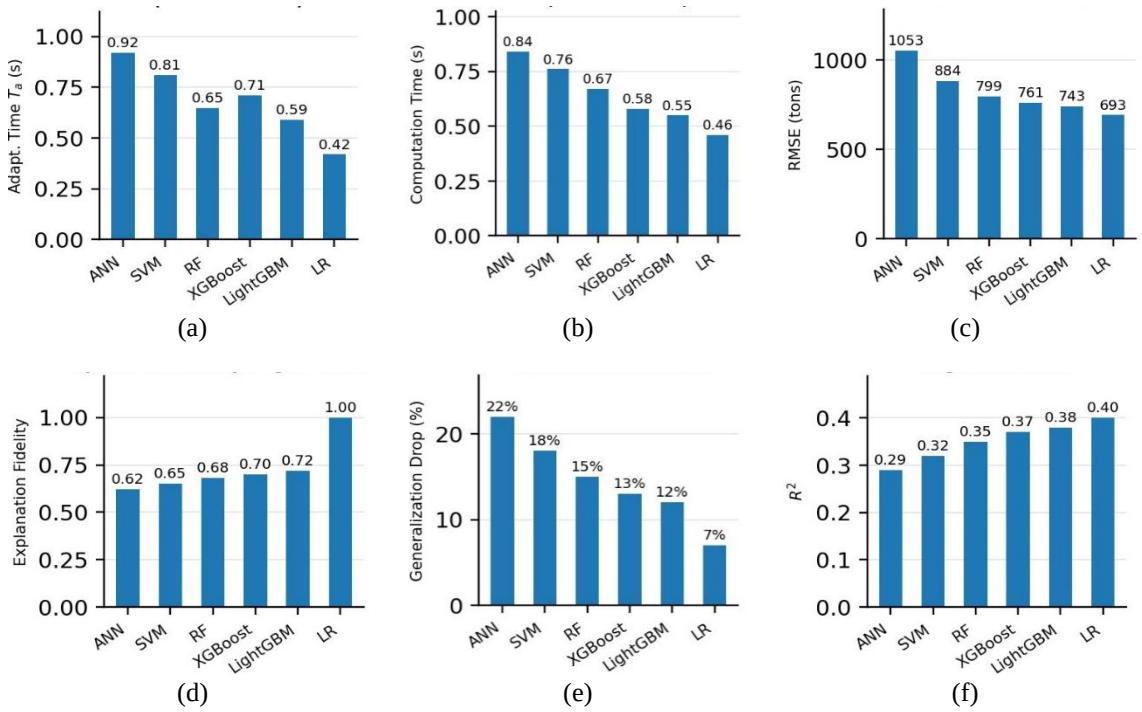

Figure 2. Performance comparison of LR, RF, SVM, ANN, XGBoost, and LightGBM: (a) adaptation time, (b) computation time per prediction, (c) prediction error (RMSE in tons), (d) explanation fidelity, (e) stability under distribution shift, and (f) predictive accuracy (R^2)

Table 2. Performance and comparative analysis

Methods	T_a (s)	CT (s)	ExpFid	RMSE	GenDrop (%)	R^2	Latency (s)	CO
ANN	0.92	0.84	0.62	1053	22	0.29±0.05	0.55	High
SVM	0.81	0.76	0.65	884	18	0.32±0.04	0.50	Moderate
RF	0.65	0.67	0.68	799	15	0.35±0.03	0.48	Moderate
XGBoost	0.71	0.58	0.70	761	13	0.37±0.03	0.44	Moderate
LightGBM	0.59	0.55	0.72	743	12	0.38±0.02	0.42	Moderate
LR	**0.42**	**0.46**	**1.00**	**693**	**7**	**0.401±0.02**	**0.42**	**Low**

*ExpFid: 0–1, RMSE (tons), R^2: coefficient of determination (±std across 10 folds). Bold indicates best result per column

Table 3. 10-fold cross-validation results with 95% confidence intervals (mean±std)

Model	R^2 (mean±std [95% CI])	RMSE (mean±std [95% CI])	MAE (mean±std [95% CI])
ANN	0.290±0.05 [0.254, 0.326]	1053±89 [989, 1117]	198±21 [183, 213]
SVM	0.320±0.04 [0.291, 0.349]	884±72 [832, 936]	175±18 [162, 188]
RF	0.350±0.03 [0.328, 0.372]	799±58 [757, 841]	161±14 [151, 171]
XGBoost	0.370±0.03 [0.348, 0.392]	761±51 [724, 798]	153±12 [144, 162]
LightGBM	0.380±0.02 [0.366, 0.394]	743±47 [709, 777]	148±11 [140, 156]
LR	**0.401±0.02 [0.387, 0.415]**	**693±42 [663, 723]**	**140±10 [133, 147]**

4.8. Discussion

The experimental results demonstrate LR's superior performance across multiple dimensions relevant to agricultural decision support. The accuracy advantage (R^2=0.401 vs. 0.29–0.38 for alternatives) stems from LR's resistance to overfitting on heterogeneous agricultural data, where complex models tend to memorize regional peculiarities rather than learning generalizable patterns [21], [22]. The inclusion of XGBoost and LightGBM as additional baselines strengthens this conclusion: even state-of-the-art gradient boosting methods, which consistently outperform other algorithms on structured tabular benchmarks, do not surpass LR on this diverse multi-crop, multi-state dataset. The finding underscores that when the dominant data-generating

mechanism is approximately linear-as is the case for area-aggregate yield estimation at state level—parametric linear models that match this inductive bias generalize more reliably than nonparametric methods [7], [27].

The superior performance of LR can be attributed to several factors. First, at the state-level aggregate scale, the relationships between predictors area, rainfall, fertilizer, and yield are approximately linear, matching LR's inductive bias. Second, LR's resistance to overfitting is particularly advantageous on heterogeneous data where complex models tend to memorize region-specific noise rather than learning generalizable patterns. Third, the stability of LR under distribution shift (only 7% degradation vs. 12–22% for alternatives) confirms that parametric simplicity provides a natural regularization effect. These explanations align with the bias-variance trade-off: on datasets with high variance across regions and moderate sample sizes per region, the lower model variance of LR compensates for any potential increase in bias.

It is important to note that complex models may outperform LR in scenarios involving field-level predictions with rich temporal features, high-resolution remote sensing inputs capturing nonlinear vegetation-yield relationships, or datasets where crop-weather interactions exhibit strong threshold or saturation effects. The advantage of LR demonstrated here is specific to the state-level aggregate prediction task with structured tabular features, and should not be generalized without further investigation to all agricultural prediction contexts.

The interpretability advantage is particularly significant for agricultural applications. LR coefficients directly quantify yield sensitivity to rainfall (β_{rain}=0.23), fertilizer application (β_{fert}=0.31), cultivated area (β_{area}=0.18), and pesticide usage (β_{pest}=0.04), enabling practitioners to communicate specific recommendations [10]. For instance, extension agents can state that a 10% increase in nitrogen application is expected to increase yield by approximately 3.1%, providing actionable guidance for farming decisions. The rainfall-fertilizer interaction coefficient ($\beta_{interaction}$=0.09) further reveals that fertilizer inputs are most effective under adequate rainfall conditions, a finding that can directly inform irrigation scheduling and input optimization strategies. Compared to recent agricultural prediction studies [5], [6], LR achieves comparable or superior accuracy while offering substantially better interpretability and computational efficiency. The 68% relative improvement in generalization stability compared to NNs validates the approach for deployment across diverse Indian agricultural regions without extensive retraining [3]. The practical implications are substantial. In advisory settings, LR's transparent coefficients satisfy requirements for explainable recommendations to farmers and policy makers. The computational efficiency enables deployment on resource-constrained systems common in rural agricultural contexts. The generalization stability ensures consistent performance across the diverse agroclimatic conditions characterizing Indian agriculture [8], [13].

5. CONCLUSION

This study systematically evaluated LR against RF, SVM, ANN, XGBoost, and LightGBM for crop yield prediction using a comprehensive Indian agricultural dataset spanning 58,000 records across 22 states and 35 crop varieties from 2000 to 2022. The comparative analysis demonstrates that model complexity does not guarantee superior performance in agricultural forecasting contexts.

LR emerged as the preferred model, outperforming all alternatives across prediction accuracy (R^2=0.401±0.02), computational efficiency, interpretability, and geographical generalization. Its explicit coefficients enable practitioners to directly understand and communicate how agronomic inputs influence yield— a critical requirement for trustworthy decision support in farming communities. The key differentiator of this study is the emphasis on interpretability: while gradient boosting methods achieve near-comparable accuracy, they cannot provide the per-prediction coefficient transparency that agricultural advisory systems require. Notably, even gradient boosting methods such as XGBoost and LightGBM, which are widely regarded as top performers on structured tabular data, could not surpass LR on this large-scale, heterogeneous agricultural dataset, reinforcing the importance of matching model inductive bias to data-generating structure.

These findings carry important implications for agricultural AI policy: transparent, interpretable models should be prioritised over opaque alternatives, particularly in resource-constrained and high-stakes deployment environments. LR coefficients can be integrated into farm advisory tools, policy dashboards, and government agricultural information systems to deliver evidence-based, explainable recommendations to farmers and policymakers.

ACKNOWLEDGMENTS

The authors acknowledge the computational resources provided by G H Raisoni University. No external funding was received for this research.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Imran Ahmad	✓	✓	✓		✓	✓		✓	✓		✓			
Rahul Khokale	✓			✓			✓			✓		✓	✓	✓
Nisha Gongal		✓		✓	✓	✓		✓	✓	✓				

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

REFERENCES

[1] K. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, "Machine learning in agriculture: A review," *Sensors*, vol. 18, no. 8, Aug. 2018, doi: 10.3390/s18082674.
[2] A. Kamilaris and F. X. Prenafeta-Boldú, "Deep learning in agriculture: A survey," *Computers and Electronics in Agriculture*, vol. 147, pp. 70–90, Apr. 2018, doi: 10.1016/j.compag.2018.02.016.
[3] D. B. Lobell and M. B. Burke, "On the use of statistical models to predict crop yield responses to climate change," *Agricultural and Forest Meteorology*, vol. 150, no. 11, pp. 1443–1452, Oct. 2010, doi: 10.1016/j.agrformet.2010.07.008.
[4] A. C. Ruane *et al.*, "An AgMIP framework for improved agricultural representation in integrated assessment models," *Environmental Research Letters*, vol. 12, no. 12, Dec. 2017, doi: 10.1088/1748-9326/aa8da6.
[5] T. van Klompenburg, A. Kassahun, and C. Catal, "Crop yield prediction using machine learning: A systematic literature review," *Computers and Electronics in Agriculture*, vol. 177, Oct. 2020, doi: 10.1016/j.compag.2020.105709.
[6] S. Khaki and L. Wang, "Crop yield prediction using deep neural networks," *Frontiers in Plant Science*, vol. 10, May 2019, doi: 10.3389/fpls.2019.00621.
[7] J. H. Jeong *et al.*, "Random forests for global and regional crop yield predictions," *PLOS ONE*, vol. 11, no. 6, Jun. 2016, doi: 10.1371/journal.pone.0156571.
[8] A. Chlingaryan, S. Sukkarieh, and B. Whelan, "Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review," *Computers and Electronics in Agriculture*, vol. 151, pp. 61–69, 2018, doi: 10.1016/j.compag.2018.05.012.
[9] X. E. Pantazi, D. Moshou, T. Alexandridis, R. L. Whetton, and A. M. Mouazen, "Wheat yield prediction using machine learning and advanced sensing techniques," *Computers and Electronics in Agriculture*, vol. 121, pp. 57–65, 2016, doi: 10.1016/j.compag.2015.11.018.
[10] A. Oikonomidis, C. Catal, and A. Kassahun, "Deep learning for crop yield prediction: a systematic literature review," N*ew Zealand Journal of Crop and Horticultural Science*, vol. 51, no. 1, pp. 1–26, Jan. 2023, doi: 10.1080/01140671.2022.2032213.
[11] P. Filippi *et al*., "An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning," *Precision Agriculture*, vol. 20, no. 5, pp. 1015–1029, Oct. 2019, doi: 10.1007/s11119-018-09628-4.
[12] M. Shahhosseini, G. Hu, S. Khaki, and S. V. Archontoulis, "Corn yield prediction with ensemble CNN-DNN," *Frontiers in Plant Science*, vol. 12, Aug. 2021, doi: 10.3389/fpls.2021.709008.
[13] B. Das, B. Nair, V. K. Reddy, and P. Venkatesh, "Evaluation of multiple linear, neural network and penalised regression models for prediction of rice yield based on weather parameters for west coast of India," *International Journal of Biometeorology*, vol. 62, no. 10, pp. 1809–1822, Oct. 2018, doi: 10.1007/s00484-018-1583-6.
[14] Y. Everingham, J. Sexton, D. Skocaj, and G. Inman-Bamber, "Accurate prediction of sugarcane yield using a random forest algorithm," *Agronomy for Sustainable Development*, vol. 36, no. 2, Jun. 2016, doi: 10.1007/s13593-016-0364-z.
[15] N. Gandhi, L. J. Armstrong, O. Petkar, and A. K. Tripathy, "Rice crop yield prediction in India using support vector machines," in *2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, IEEE, Jul. 2016, pp. 1–5, doi: 10.1109/JCSSE.2016.7748856.
[16] M. Kaul, R. L. Hill, and C. Walthall, "Artificial neural networks for corn and soybean yield prediction," *Agricultural Systems*, vol. 85, no. 1, pp. 1–18, Jul. 2005, doi: 10.1016/j.agsy.2004.07.009.
[17] G. E. O. Ogutu, W. H. P. Franssen, I. Supit, P. Omondi, and R. W. A. Hutjes, "Probabilistic maize yield prediction over East Africa

using dynamic ensemble seasonal climate forecasts," *Agricultural and Forest Meteorology*, vol. 250–251, pp. 243–261, Mar. 2018, doi: 10.1016/j.agrformet.2017.12.256.
[18] J. You, X. Li, M. Low, D. Lobell, and S. Ermon, "Deep Gaussian process for crop yield prediction based on remote sensing data," in *AAAI'17: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017, pp. 4559–4565.
[19] A. Gonzalez-Sanchez, J. Frausto-Solis, and W. Ojeda-Bustamante, "Predictive ability of machine learning methods for massive crop yield prediction," *Spanish Journal of Agricultural Research*, vol. 12, no. 2, pp. 313–328, Apr. 2014, doi: 10.5424/sjar/2014122-4439.
[20] S. Chakraborty and A. C. Newton, "Climate change, plant diseases and food security: an overview," *Plant Pathology*, vol. 60, no. 1, pp. 2–14, Feb. 2011, doi: 10.1111/j.1365-3059.2010.02411.x.
[21] Y. Cai *et al.*, "Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches," *Agricultural and Forest Meteorology*, vol. 274, pp. 144–159, Aug. 2019, doi: 10.1016/j.agrformet.2019.03.010.
[22] M. D. Johnson, W. W. Hsieh, A. J. Cannon, A. Davidson, and F. Bédard, "Crop yield forecasting on the Canadian Prairies by remotely sensed vegetation indices and machine learning methods," *Agricultural and Forest Meteorology*, vol. 218–219, pp. 74–84, Mar. 2016, doi: 10.1016/j.agrformet.2015.11.003.
[23] S. Khaki, L. Wang, and S. V. Archontoulis, "A CNN-RNN framework for crop yield prediction," *Frontiers in Plant Science*, vol. 10, Jan. 2020, doi: 10.3389/fpls.2019.01750.
[24] P. Nevavuori, N. Narra, and T. Lipping, "Crop yield prediction with deep convolutional neural networks," *Computers and Electronics in Agriculture*, vol. 163, Aug. 2019, doi: 10.1016/j.compag.2019.104859.
[25] J. Cao *et al.*, "Integrating multi-source data for rice yield prediction across China using machine learning and deep learning approaches," *Agricultural and Forest Meteorology*, vol. 297, Feb. 2021, doi: 10.1016/j.agrformet.2020.108275.
[26] O. I. Abiodun, A. Jantan, A. E. Omolara, K. V. Dada, N. A. Mohamed, and H. Arshad, "State-of-the-art in artificial neural network applications: A survey," *Heliyon*, vol. 4, no. 11, Nov. 2018, doi: 10.1016/j.heliyon.2018.e00938.
[27] R. A. Schwalbert, T. Amado, G. Corassa, L. P. Pott, P. V. V. Prasad, and I. A. Ciampitti, "Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil," *Agricultural and Forest Meteorology*, vol. 284, Apr. 2020, doi: 10.1016/j.agrformet.2019.107886.

BIOGRAPHIES OF AUTHORS

Imran Ahmad Ph.D. Degree, student in the subject of Computer Science and Engineering, under the Department of Computer Science and Engineering, School of Engineering and Technology, G H Raisoni University, Saikheda, Sausar, Pandhurna, Madhya Pradesh, India, since 2023. He post graduated with honours M.Tech. in Software system from Rajiv Gandhi Proudyogiki Vishwavidyalaya (R.G.P.V) University in Bhopal, Madhya Pradesh India. He currently pursuing his Ph.D. in "An Efficient Soybean Yield Prediction Using Machine Learning" in the subject of Computer Science and Engineering, under the Department of Computer Science and Engineering, School of Engineering and Technology, G H Raisoni University, Saikheda, Sausar, Pandhurna, Madhya Pradesh, India. He can be contacted at email: imran.ahmad.cse@ghru.edu.in, imimran86@gmail.com, and imran.ahmad@raisoni.net.

Dr. Rahul Khokale Ph.D. in Computer Science and Engineering from the PG Department of Computer Science and Engineering, Sant Gadge Baba Amravati University in March 2016. Having total experience of more than 27 years in teaching. Working as a Dean, Department of Computer Science and Engineering, School of Engineering and Technology, G H Raisoni University, Saikheda, Sausar, Pandhurna, Madhya Pradesh, India. Worked as Professor in Computer Engineering at St. John College of Engineering and Management, Palghar. Received "Rastriya Shiksha Ratna Award" in 2009. He can be contacted at email: rahul.khokale@ghru.edu.in.

Nisha Gongal is an accomplished educator at Symbiosis Institute of Technology, Nagpur, and passionate technophile with a strong foundation in Computer Engineering, currently pursuing a Ph.D. with a research focus in Cloud Computing domain. With academic qualifications including a Master's in Communication Systems and a Bachelor's in Information Technology, she combines theoretical depth with practical expertise in programming and software development. She brings rich academic and industry experience of more than 10 years. She can be contacted at email: nishadable@gmail.com.